AD_____

Award Number: W81XWH-08-1-0402

TITLE: Exploring the Pathogenic and Therapeutic Implications of Aberrant Splicing in Breast Cancer

PRINCIPAL INVESTIGATOR: Dr. William Foulkes

CONTRACTING ORGANIZATION: Sir Mortimer B. Davis- Jewish General Hospital
Montreal H3T 1E2

REPORT DATE: July 2011

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-07-2011	Final	01 JUL 2008 - 30 JUN 2011

4. TITLE AND SUBTITLE

Exploring the Pathogenic and Therapeutic Implications of Aberrant Splicing in Breast Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER

W81XWH-08-1-0402

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Dr. William Foulkes

E-Mail: william.foulkes@mcgill.ca

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Sir Mortimer B. Davis- Jewish General Hospital
Montreal H3T 1E2

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this proposal, we set out to a) systematically monitor splicing variant profiles in breast cancer susceptibility genes and b) explore the role of alternative splicing in breast chemotherapy using a global strategy. In doing so, we hope to identify and validate candidate splicing variants involved in tumorigenesis. Although the original the molecular barcode strategy was not implemented, we have made significant findings in two areas – 1) we have identified over 1400 new splice variants and have shown that they are more prevalent in BRCA1-related breast cancer and 3) we have identified 3 novel fusion proteins. Perhaps most importantly, we have trained two young bioinformaticians, who both have bright futures in cancer research.

15. SUBJECT TERMS

Breast cancer, splicing, single molecule analysis, RNA sequencing, BRCA2, Variants of Unknown Significance, Functional assays

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	21	19b. TELEPHONE NUMBER *(include area code)*

Table of contents

W81XWH-08-1-0402

PRINCIPAL INVESTIGATORS: William D Foulkes MB PhD, Jun Zhu PhD

CO-INVESTIGATOR: Jacek Majewski, PhD

Introduction

In this proposal, we aimed to a) systematically monitor splicing variant profiles in breast cancer susceptibility genes and b) explore the role of alternative splicing in breast chemotherapy using a global strategy. In doing so, we hoped to identify and validate candidate splicing variants involved in tumorigenesis using polony digital exon-profiling and functional assays. This annual report summarizes our progress in the final year of this award (we were granted a no-cost extension).

Body

Verbatim – from S.O.W

Here I describe our results with items 1-3 from the SOW here, results for items 4-6 will follow on page 16

Objective 1: To profile splicing patterns of 100 breast cancer-related genes in 30 normal and tumor breast cell lines

Objective 2: To profile splicing patterns of 100 cancer-related genes in lymphocytes and breast tissues

Objective 3: Data analysis and validation to determine the role of aberrant splicing in breast tumourgenesis

These objectives can be summarized in terms of the expected Results/Deliverables from Year 1: 1) Optimizing the method of bar-code PCR-sequencing for studying alternative splicing at candidate loci. 2) Identifying splicing variants associated with increased breast cancer susceptibility derived from the candidate loci. Year 2 focuses more on drug sensitivity, and how splicing patterns may determine the response to certain chemotherapeutic agents.

We attempted to optimize the method of bar-code PCR-sequencing for studying alternative splicing at candidate loci. This part of the project is the responsibility of Dr. Zhu. As stated in the second annual report, we used a NimbleGen Sequence-Capture array that can selectively enrich genes of interest (3-20 Mbps) for deep sequencing analysis. The captured fragments were then subjected to 454 pyrosequencing to obtain a longer read length (~400bp), which enables better identification of alternative splicing and mutation detection. The work was completed by Ting Ni (post-doctoral fellow, funded by this award when based at Duke University, now at the National Institutes of Health).

We have identified many promising splice site mutations, indels and SNPs that we still investigating, but overall, it does not look as if we will be able to use these data to find better markers for successful dasatinib treatment of breast cancer patients. However, our collaborator Dr. Raquel Aloyz has been working on the cell biology of the differences between these two lines, and she has used partly used

our data to inform her recent study of dasatinib in chronic lymphocytic leukemia (Amrein L, Soulières D, Johnston JB, Aloyz R, p53 and autophagy contribute to dasatinib resistance in primary CLL lymphocytes. Leuk Res. 2011 Jan;35(1):99-102).

We decided to focus all of our energies on the one strategy that has emerged as the best approach to a global understanding of splicing. We have characterized alternative splicing events by sequencing the RNA of breast cancer cell lines and one breast tumor directly. Because the pace of technological innovation is substantially faster than the granting process, our original proposal to analyze the splicing patterns of 100 candidate breast cancer genes (using the barcode and/or capture array strategy) has been augmented by state-of-the art whole transcriptome sequencing or RNA-Seq. In other words, in this part of the project, we have analyzed not the 100 "top" candidate breast cancer susceptibility genes, but instead all ~24,000 genes. This project has been led by PI Foulkes and co-PI Majewski. We have used Illumina/Solexa RNA sequencing to generate millions of raw reads from the tumor and control lines indicated in Table 1 below. These reads were mapped to the transcriptome using various alignment tools including BWA, TopHat and ELAND and were visualized using the Integrated Genome Viewer from the Broad Institute. In the single-end sequencing strategy, a single read is generated per transcriptome fragment. On the other hand, in paired-end strategy, the fragment is sequenced from both ends, creating a pair of reads with an expected distance apart.
We have chosen RNA-Seq for the following reasons –

- It combines the accuracy and sensitivity of Sanger sequencing with the throughput of microarrays
    - It characterizes the transcriptome at a single-base resolution, producing tens of millions of short read sequences per sample
- It measures gene and isoform expression levels and provides structural information
- It is able to reveal:
    - SNPs and indels
    - Allelic expression
    - Gene fusion and novel gene isoform transcripts
    - Gene expression

We have chosen to focus on the *BRCA1* transcriptome because we have access to quality samples and because we felt that in a highly competitive field, we needed to find a niche. At the current time, whole transcriptome sequencing is too expensive to allow us to stick to our original plan, so instead of studying 100 genes in many samples, we have opted to study all genes in few samples. As prices fall, we will be able to study more samples.

Here, we intend to -

- Compare *BRCA1*-deficient breast cancer samples to controls
- Characterize the *BRCA1* breast cancer transcriptome via next-generation RNA sequencing (RNA-seq) by identifying:
    - fusion proteins – a type of spliced protein where exons from two different genes are spliced together to make a new protein, with potentially new functions.
    - new splice isoforms - the main objective of the current proposal
      SNPs and indels - this is not the main thrust of the current proposal, but since we have these data, we will be able to study these at a later date

Positive results can provide new biomarkers and detect new proteins or pathways involved in *BRCA1*-related breast cancer formation

Table 1: Cell lines and tumors used in this study.

Fourteen breast cancer samples were sequenced in this study using either a single end (SE) or paired end (PE) approach. The first seven samples have known germline BRCA1 mutations (based on RefSeq accession NM_007294) and the last seven samples are the controls (no germline BRCA1 mutations).

Sample	Description	*BRCA1* germline mutation	Read length (bp)	Lanes	Sequencing center
HCC1937	Breast cancer cell line	5266dupC	50 SE	3	Illumina
SUM149PT	Breast cancer cell line	2288delT	36 PE	3	GQIC
HCC3153	Breast cancer cell line	68_69delAG	36 PE	3	GQIC
SUM1315	Breast cancer cell line	943ins10	36 PE	3	GQIC
T92	Primary breast tumor	5266dupC	54 PE	1	GQIC
T50	Primary breast tumor	4328G>A	54 PE	1	GQIC
T160	Primary breast tumor	5521T>G	54 PE	1	GQIC
SEC1	Primary breast tumor	-	54 PE	2	ICR
SEC2	Primary breast tumor	-	54 PE	2	ICR
MBC647	Breast metasplastic carcinoma cell line	-	54 PE	2	ICR
MPC7105	Breast micropapillary carcinoma cell line	-	54 PE	2	ICR
MPC298	Breast micropapillary carcinoma cell line	-	54 PE	2	ICR
MPC600	Breast micropapillary carcinoma cell line	-	54 PE	2	ICR
MPC960	Breast micropapillary carcinoma cell line	-	54 PE	2	ICR

The millions of RNA-Seq reads are aligned to splice junction libraries, and we use these reads to quantify and characterize alternative splicing events. However, one of our main interests was to identify novel splice isoforms (not present in most reference splice libraries). To this end, we have pursued three approaches: 1) Mapping sequencing reads to exhaustive splice junction libraries representing all possible splice junctions *within* each gene; 2) creating heuristic algorithms for mapping reads to an exhaustive library joining all exons *across* all genes in the human genome 3) monitoring excess reads that map to normally intronic sequences – this is meant to detect novel intron retention events;. The second approach will allow us to detect all known and novel alternative splicing events, trans-splicing (across genes), as well gene fusion events which are known to occur in many tumors; . In other words, we are not just looking for known splicing events, but all possible combinations of exons within a given gene as well as between any two genes in the genome. Given the high rate of pervasive splicing in the human genome and the ability of RNA-Seq to detect these variants, finding transcript isoforms which

have low expression but large effect sizes (such as dominant negative splice variants) is daunting and unrealistic using sequencing data alone. Therefore, we focused on potentially abnormal, novel splice variants (defined in this manuscript as splice junctions absent from the RefSeq and UCSC databases) that have high relative expression levels in the majority of the *BRCA1*-related breast cancers, indicative of a driver variant conferring selective advantage to this subtype of breast cancer.

A) Splicing results

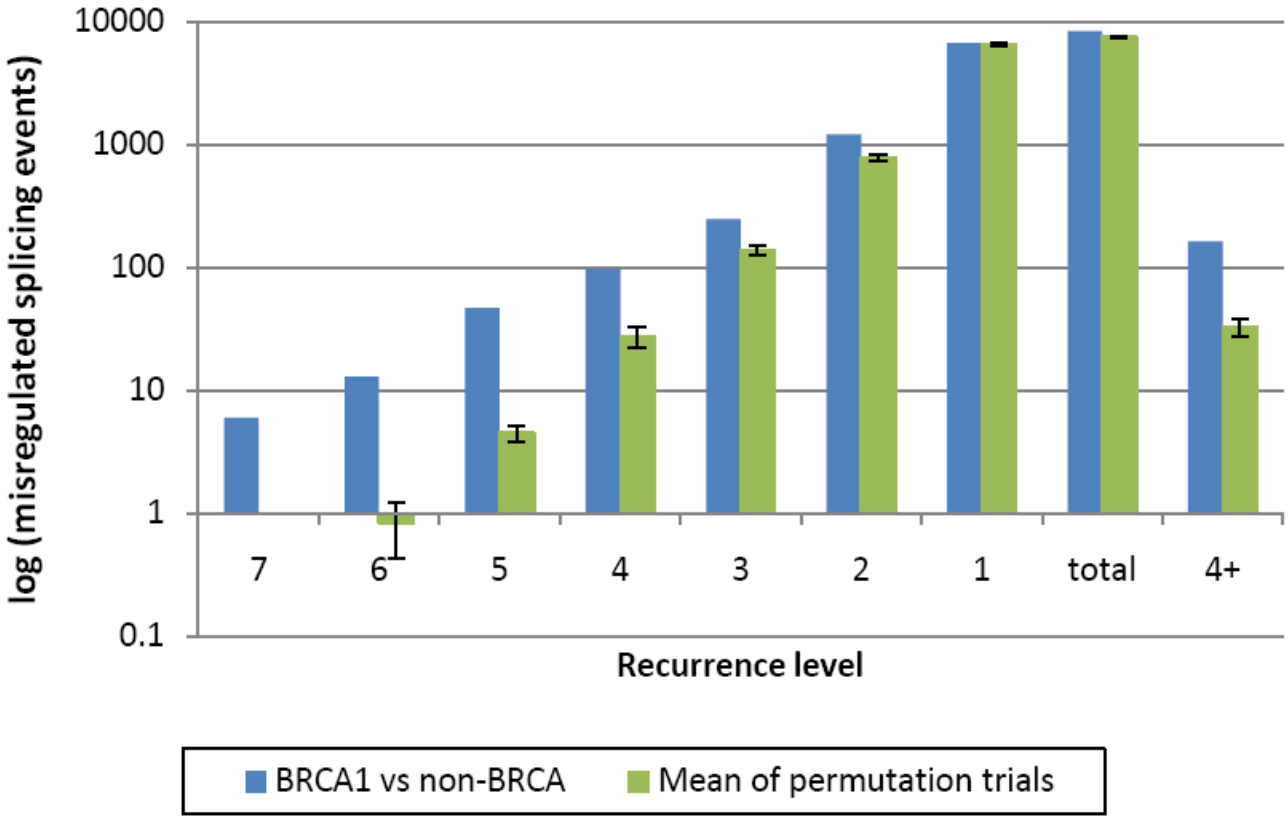
First, we present our results from the splicing analysis of the cell lines/tumor in Table 1 above. Note the very large number (>1400) of novel junctions identified in at least two of the breast cancer samples yet absent or with very low expression in the control samples (Figure 1). We ignore events which were found in a single breast cancer sample as our aim is to identify novel splice variants which may characterize *BRCA1* breast cancers and as such must be recurrent. These results are summarized in Table 2

Table 1. The number of abnormal splicing events identified using variable thresholds. This table shows the number of abnormal splicing events identified with our pipeline using variable thresholds for increase in expression and recurrence levels (the number of *BRCA1*-related breast cancers in which abnormal junctions pass all pipeline filters). The relative expression refers to the fold increase in junction ratio of a *BRCA1*-related breast cancer sample compared to the average junction ratio for the controls. For example, two abnormal junctions had junction ratios at least 1.5 times greater in all all 7 *BRCA1* samples compared to the controls. Our pipeline thresholds were selected in a pragmatic manner so that the number of candidates to evaluate would be manageable while maximizing the potential biological impact of the candidate variants.

		Increase in expression levels			
		1.5X	**2X**	**5X**	**10X**
Recurrence level	**1 of 7**	5075	4337	2599	2064
	2 of 7	955	649	279	192
	3 of 7	248	122	30	24
	4 of 7	84	44	10	2
	5 of 7	29	14	0	0
	6 of 7	16	6	0	0
	7 of 7	2	0	0	0
	Total	6409	5172	2918	2282

It can be observed that as the stringency increases, the number of novel splice junctions decreases very steeply. Arbitrarily, we thought that a 2-fold difference in splicing compared with controls, in 6 out of 7 BRCA1- associated tumors or cell lines was a good cut-off, and thus we had 14 such novel junctions.

Figure 1: BRCA1 tumors have more recurrent novel splicing junctions than do non-BRCA1 samples
A higher number of junctions were identified in the *BRCA1* samples than expected by chance. We applied our Abnormal Alternative Splicing Pipeline to six randomly permutated case-control sets and compared the mean number of junctions identified at each recurrence level to the number found in our true case-control experiment (i.e. “BRCA1 vs non-BRCA”). Up- and down- regulated abnormal splice variants are shown. Error bars indicate the standard errors of the six random trials.



Another way to look at the data is shown in Table 3. He we look at highly discordant expression levels, and set the recurrence level more liberally, at 5 of 7 tumors. Here we identified 20 novel splicing events.

Table 2 (next page). Abnormal splice junctions with the most discordant expression values between *BRCA1* and non-*BRCA1* breast cancers with recurrence levels (RL) greater than or equal to 5. Unless otherwise stated, the alternative splicing event (ASE) does not cause a frameshift. For each ASE, we determined whether the new isoform had putative effects on a protein domain, on cancer progression or on gene expression. Protein domain information was obtained from the Human Protein Reference Database (HPRD). If no domains were found in HPRD, NDF is noted (No Domain Found). Gene functions were found in the UCSC database unless otherwise noted.

Gene	RL	Type of event	Protein sequence/ protein domain affected?	Gene involved in cancer?	Gene involved in expression control?
USMG5	6	Novel exon in 5'UTR	N / N	N	N
USMG5	6	Novel exon in 5'UTR	N / N	N	N
SUMO2	6	Alternative initation	Y / N	N	Y
TMUB2	6	Alternate 5' SS in 5'UTR exon	N / N	N	N

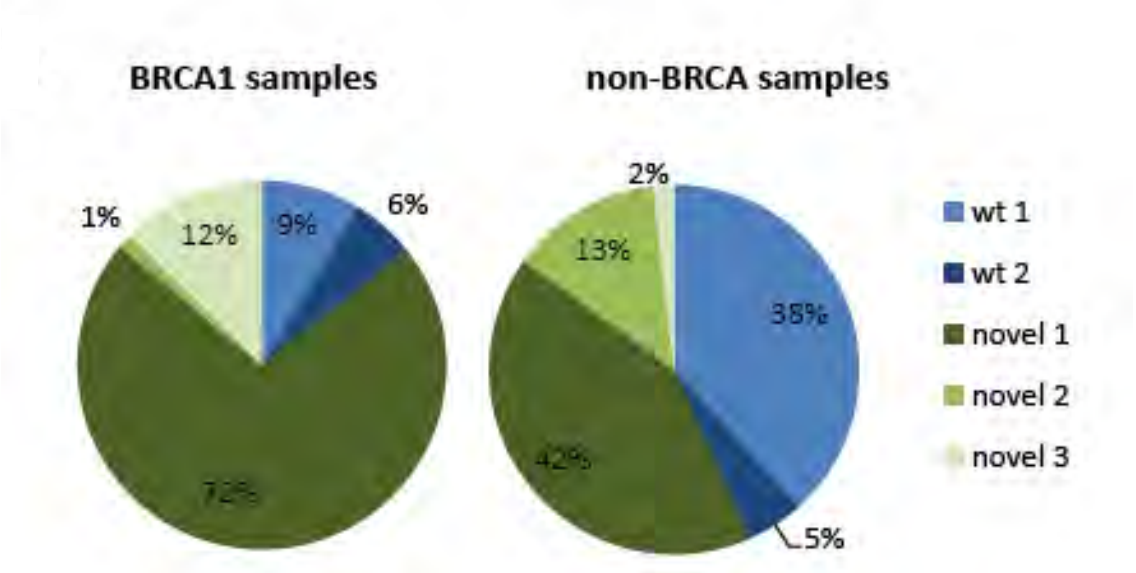
TMEM167A	6	Cassette exon and frameshift	Y/Y- transmembrane domain	N	N
SNRPG	6	Cassette exon and frameshift	Y / Y- Sm domain of snRNPs proteins	N	N
ADRM1	5	Cassette exon	Y / N	Y- Knockdown suppresses cell proliferation in colorectal cancer via apoptosis and cell cycle arrest [39]. Up-regulated in various solid tumors	Y- involved in the ubiquination pathway
EMP2	5	Novel exon in 5'UTR	N / N	Y- knockdown induces apoptosis and slowed cell growth in cancer cells [40]	Y- associated with GO process "Cell proliferation"
GAS5	5	Exon skipping event in 5'UTR	N / N	Y- can sensitize cells to apoptosis and is down-regulated in breast cancer[41]	Y- can sensitize cells to apoptosis [41]
PCBP2	5	Novel terminal exon, 12 amino acids skipped	Y / Y- K Homology domain	Y- Overexpression induces apoptosis in cancerous cell lines [42]	Y- transcriptional regulator of gene expression [43]
SUPT4H1	5	Cassette exon and frameshift	Y / NDF	N	Y- part of a complex which regulates mRNA processing and transcription elongation
UQCRB	5	Novel coding exon and frameshift	Y / NDF	N	N
LOC387647	5	Exon skipped in 3'UTR, 8 amino acids skipped	N / N	N	N
TMEM141	5	Cassette exon and frameshift	Y/Y- transmembrane domain	N	N
FAM113A	5	Cassette exon	Y / NDF	N	N
SNHG8	5	Alternate 5'SS in 5'UTR exon	N / N	N	N
PTP4A2	5	Cassette exon	Y / Y- Tyrosine phosphotase domain	Y- Overexpression in mammalian cells conferred a transformed phenotype, which suggested its role in tumorigenesis[44]	N
NFIC	5	Cassette exon and frameshift	Y / NDF	N	N
CCDC58	5	Alternate 5' SS and frameshift	Y / NDF	N	N

UBE2S	5	Novel coding exon	Y / N	Y- Overexpression increases tumor cell proliferation, invasion, and metastasis by destabilizing VHL [45]	Y- involved in the ubiquination pathway

Figure 2. **An abnormal splicing event in *PTP4A2*is more abundant in*BRCA1*-related breast cancers than in other breast cancers.**
Transcripts skipping exon 3 (starred) have a higher proportion of expression than other transcripts from this locus in *BRCA1* breast cancers. **A**- The different transcripts either including or excluding exon 3. Each transcript may actually represent several transcripts depending on exon usage up- and down-stream of exon 3. "Novel 1" is the transcript identified by our analysis and is uniquely identified by junction 8. **B**- The average proportion of each isoform from C in *BRCA1* breast cancers and non-*BRCA* breast cancers, colour-coded as in B. The number of reads supporting a unique junction from each transcript was used to compare the abundance of each transcript

Figure 2A:

Figure 2B:



We have also looked at all splicing events, not just novel junctions, and have identified several pathways that may be implicated in BRCA1-related breast cancer.

Summary of findings from the splicing project

BRCA1-related breast cancers are generally of high histological grade, associated with adverse prognosis and often occur with other primary and metastatic tumors. Perturbation of various cellular processes underlies the aggressive phenotype of these tumors. We postulated that faulty RNA splicing may also contribute to BRCA1-related breast cancers. In particular, we were interested in investigating if aberrant expression of specific isoforms may be directly involved in the BRCA1 cancer phenotypes. RNA-Seq proved to be an effective tool in discovering abnormal splice variants in a global and unbiased manner. However, most abnormal transcripts identified were also present in non-BRCA1 breast cancers or healthy tissues, suggesting that the major gene annotation databases underestimate the transcriptional diversity at many gene loci. Of note, some variants identified do have pathogenic potential and follow-up functional studies are warranted. In the future, the use of longer read lengths will help in elucidating how multiple exons are combined, rather than only two exons, to gain a greater understanding of the structure of full transcripts present in the samples. Third generation sequencing, or single-cell sequencing, will also be invaluable in such studies by eliminating issues of cellular heterogeneity and mixed cell populations, which may cause under-estimation of expression values.

This work shows that although BRCA1-related breast cancers appear to share a pattern of mRNA splicing dysregulation, we did not find evidence for any highly expressed and highly recurrent abnormal transcripts which could characterize this type of breast cancer. Because of the massive amounts of data produced by RNA-Seq and the high rate of missplicing in the genome, we restricted our search to variants with high expression and specificity patterns. However, using these strategies we cannot confidently conclude that abnormally spliced transcripts contribute to the BRCA1 phenotype. It is nevertheless possible that rare, pathogenic transcripts which would affect a small subset of BRCA1 cancers do exist Although we did not find evidence that genetic regulation at the mRNA level confers strong phenotypic variation in BRCA1-related breast cancers, future studies with larger sample sizes will be better aimed to answer what role is played by alternative and abnormal splicing in these cancers.

B) Fusion protein project – rationale and results

Gene fusions are the result of chromosomal alterations involving two genes. These chimeras may have severe phenotypic effects, such as the well-studied *BCR-ABL1* fusion protein implicated in chronic myelogenous leukemia (Shtivelman et al., 1985) and *TMPRSS2-ERG* found in many cases of prostate cancer (Tomlins et al., 2005). New efforts using high-throughput sequencing have resulted in new discoveries of gene fusions. This has prompted interest in determining whether these chromosomal aberrations may be specific to cancer and if they are, may theoretically serve as an ideal diagnostic and therapeutic target (Prensner and Chinnaiyan, 2009).

<u>Gene fusions - examples and summary of our results</u>

We have devised two strategies to identify gene fusions using either single-end or paired-end data, following a similar approach described by Maher et al. (2009). In both cases, we downloaded from the UCSC Genome Database (hg18 assembly) the set of exon genomic sequences from all mature mRNA RefSeq transcripts. As the mRNA of gene fusions would typically involve the fusion of exon sequences from two different genes, we retained only the boundary sequences of each exon (i.e. 49 bp sequence from the left boundary of an exon and 49 bp sequence from the right boundary of an exon, for 50bp reads), in order to identify reads that may span such exon-exon boundaries.

The basic principle for single-end data is described below using the HCC1937 dataset as an example. Of

a total of 40 million 50 bp reads sequenced from this cell line, approximately 1.5 million reads were determined to be non-mapping after analyzing the data with BWA. Non-mapping reads are those which did not have a unique alignment to the reference genome. We limited our single-end gene fusion analysis to only these reads as they may potentially characterize breakpoints of gene fusions not found in the reference genome. Blat (Kent, 2002) was used to align the non-mapping reads against the list of exon boundary sequences. Using a set of in-house scripts written in Python, we filtered the results for alignments such that a read was partially aligning to an exon boundary. Of these partial alignments, we further identified whether its remaining unaligned sequence aligned to another exon region either from the Blat analysis or by simple string matching techniques. This methodology will identify gene fusions with breakpoints located within introns but not necessarily those with breakpoints in exons. However, since exons make up only 5% of the genome, we assume that the breakpoint events occur within the intron. Moreover, if a breakpoint does occur within an exon, the coding frame would have to be maintained in order for the fusion to be expressed.

For paired-end reads, we took advantage of the additional distance information. As previously mentioned, each paired reads should map at an expected distance from each other in the transcriptome. Thus, potential gene fusion events can be inferred by paired reads which map on two different mRNA corresponding to two different genes (see Figures 3-6). The genes may be located on the same chromosome (implying an intrachromosomal rearrangement) or different chromosome (interchromosomal). Thus, candidate gene fusion events are nominated based on satisfying two criteria. First, we look for a sufficient amount of supporting paired reads that map unusually in two specific loci. Secondly, we then generate potential exon-exon junction sequences joining the two genes and search for additional individual reads that map across the two exons. The latter step is analogous to the previously described single-end approach and provides further supporting evidence for the candidate gene fusion.

We first implemented the single-end procedure and tested it on HCC1937, where we successfully identified the *NFIA-EHF* gene fusion (Figure 3). This fusion was previously identified by whole-genome DNA sequencing and validated by RT-PCR and FISH in a study by Stephens et al. (*Nature*, 2009). Hence, we demonstrated as proof of principle that we can independently identify the same gene fusion using RNA-Seq. We have now studied all the lines/tumors in table 1 and identified three additional novel gene fusions (Figures 4-6).

B) Gene fusions detected

Here we show diagrammatically the four fusions transcripts we have detected. One was previously known, and is a positive control, the others are novel. We show one fusion transcript per page.

Figure 3: RNA-Seq evidence of previously described gene fusions

We first tested our SE approach on (A) the cell line HCC1937 that harbors the fusion NFIA-EHF that is formed by a translocation between chromosomes 1 and 11. Single-end reads are shown to map across the breakpoint between exon 2 of NFIA and exon 5 of EHF, as illustrated in the schematic. Next, we tested our PE approach on (B) two primary tumors that contain ETV6-NTRK3. Results from the sample 419184 are shown. Paired reads (indicated by two solid lines joined by a dotted line) as well as single reads (red lines) are shown to map across the breakpoint between exon 5 of ETV6 and exon 14 of NTRK3.

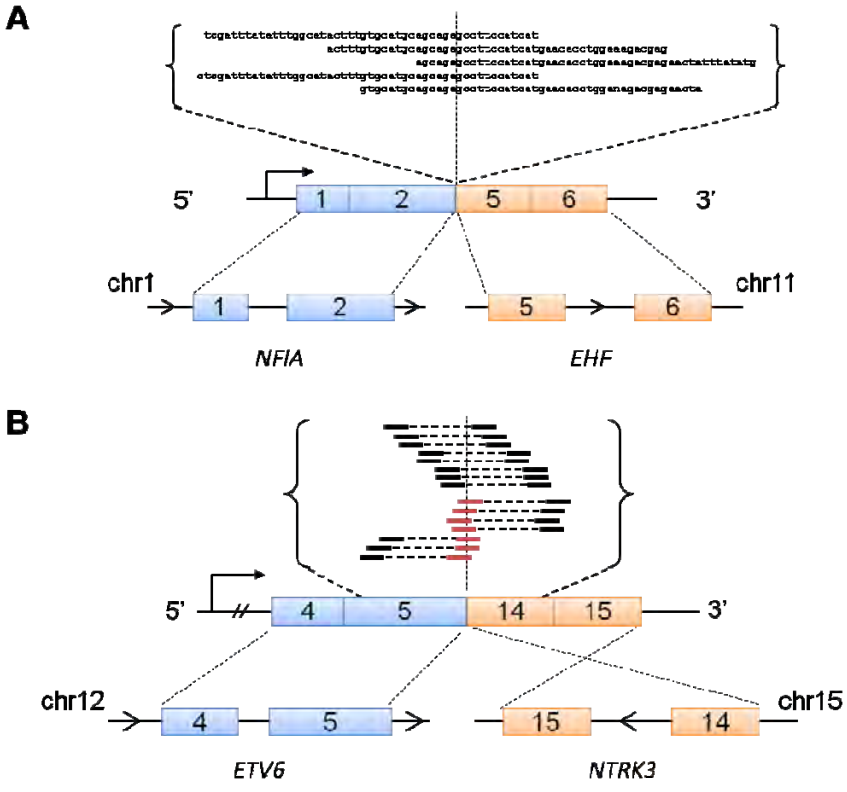


Figure 4 - RNA-Seq and Sanger sequencing evidence of WWC1-ADRBK2 gene fusion
We identified an in-frame gene fusion transcript in HCC1315: (A) Schematic of the predicted gene fusion illustrating paired-end reads that flank the breakpoint between exon 19 of WWC1 and exon 10 of ADRBK2. Reads are indicated by black solid lines. Paired reads are indicated by the dotted line joining two reads. Reads that span across the breakpoint are highlighted by red solid lines; and (B) the exon-exon breakpoint junction was tested using PCR and Sanger sequencing, which confirmed the sequence of the breakpoint junction.

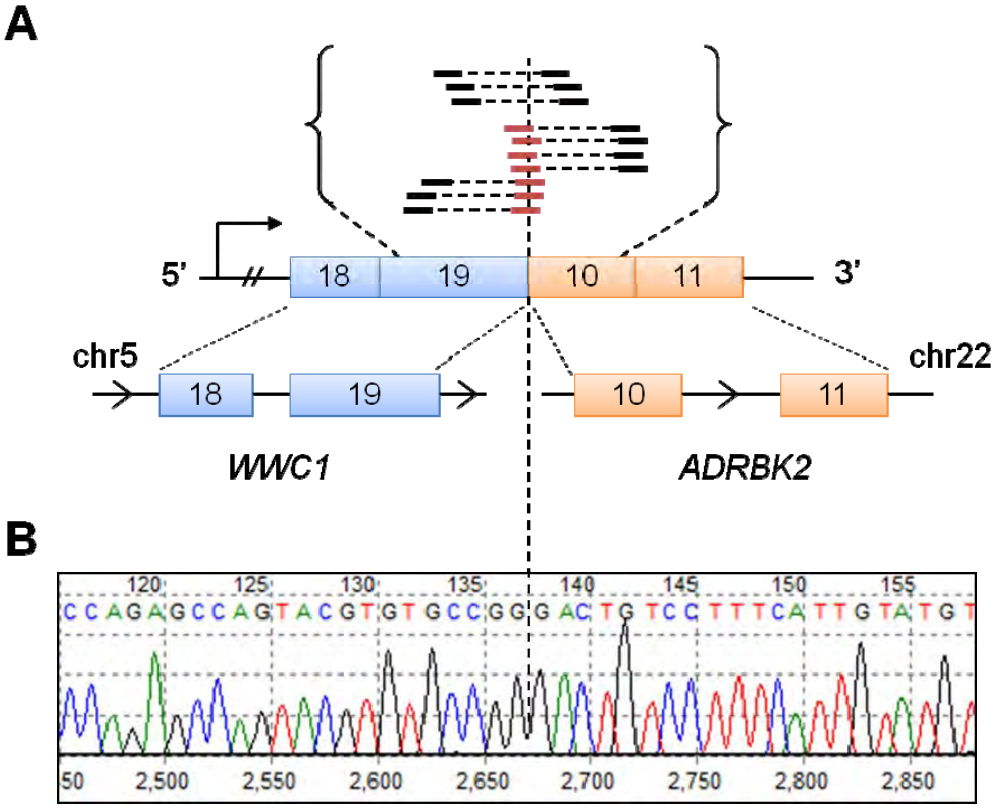


Figure 5 (next page) - RNA-Seq and Sanger sequencing of *ADNP-C20orf132*
We identified another in-frame gene fusion in a primary tumor that was present as two transcript isoforms: (A) schematic of the first predicted gene fusion isoform illustrating RNA-Seq evidence that support the fusion between exon 1 of *ADNP* and exon 17 of *C20orf132*; (B) Sanger sequencing of the fusion junction of the first isoform; (C) schematic of the second predicted isoform in which exon 2 of *ADNP* is fused with exon 17 of *C20orf132*; and (D) Sanger sequencing of the fusion junction of the

second isoform. Reads are indicated by black solid lines. Paired reads are indicated by the dotted line joining two reads. Reads that span across the fusion junction are highlighted by red solid lines.

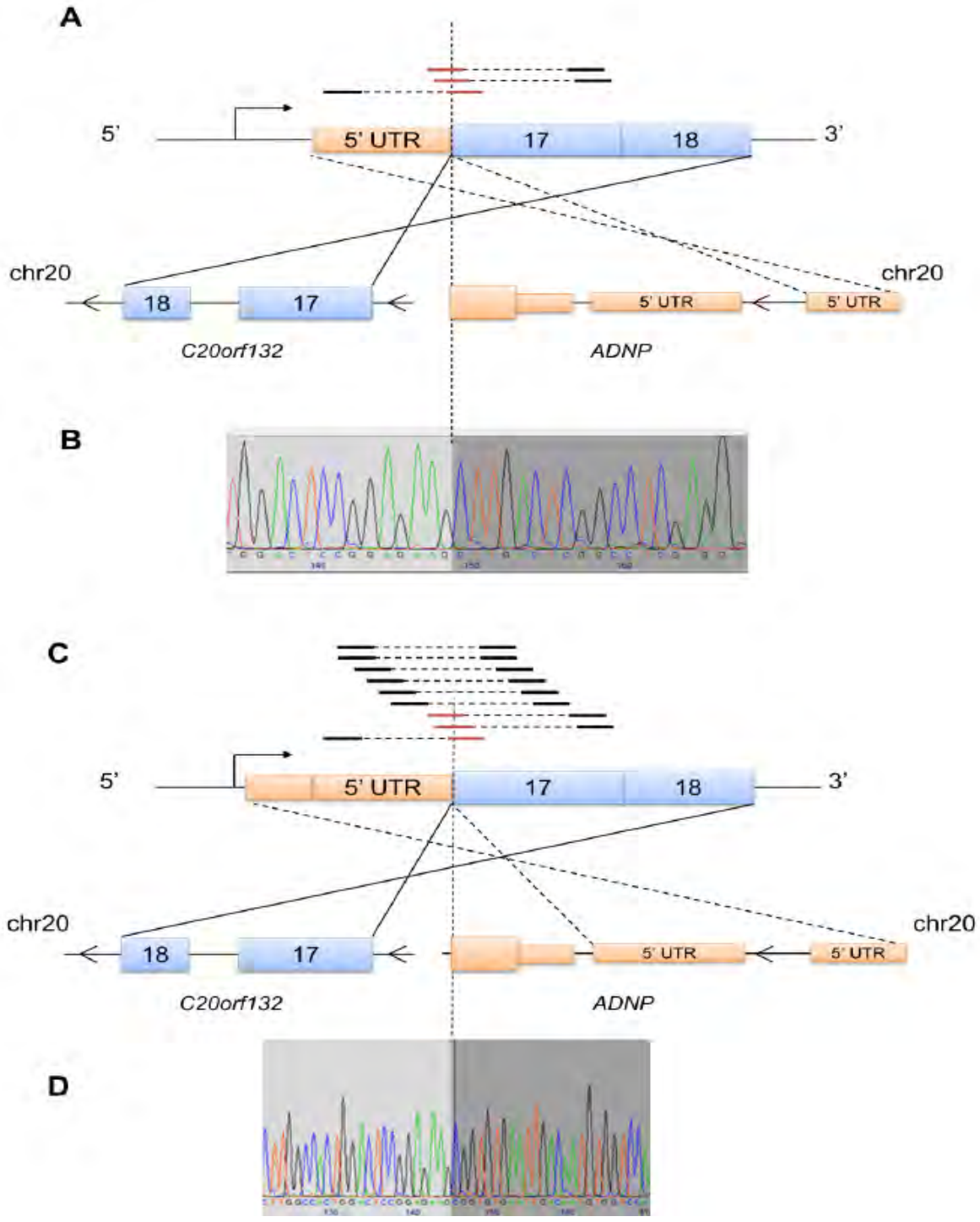
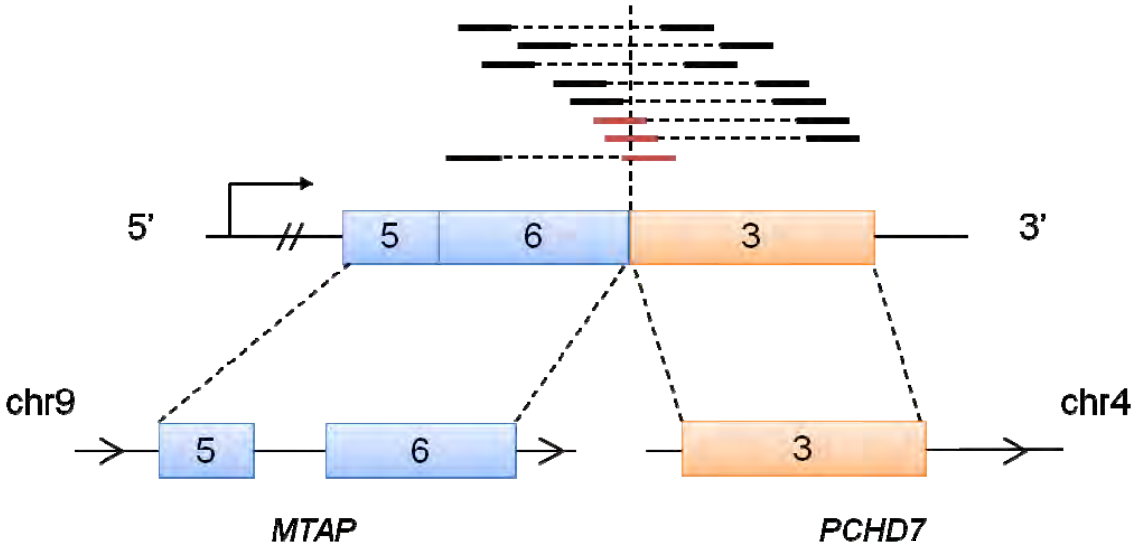


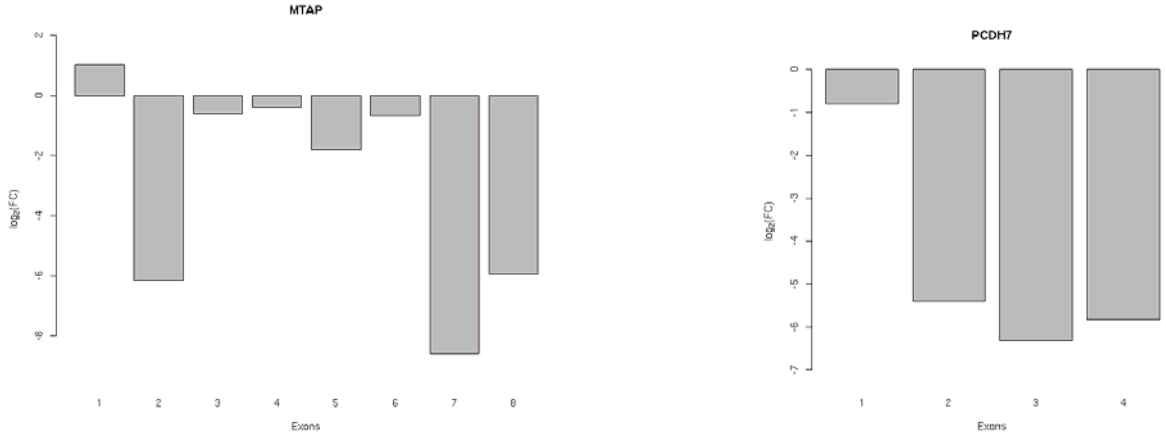
Figure 6: Novel fusion protein identified in SUM149.

A: Illustration of paired-end reads that flank the breakpoint between exon 6 of MTAP and exon 3 of

PCDH7. Reads are indicated by black solid lines. Paired reads are indicated by the dotted line joining two reads. Reads the span across the breakpoint is highlighted by red solid lines.



B: Expression plots of *MTAP* and *PCDH7* as measured by the $\log_2$ FC between the RPKM values of each exon in SUM149PT versus the average of all other *MTAP-PCDH7*-negative samples.



Summary of findings from the fusion transcript project

In this project, we have used both single-end and paired-end sequencing strategies to discover gene fusions in breast cancer transcriptomes with BRCA1 mutations. RNA-Seq with paired-end reads is an effective tool for transcriptome profiling of gene fusions. This is because the task of finding misaligned read pairs is systematically easier than locating single-end reads that partially align to the exons of different genes. Based on our analysis, our findings suggest that while gene fusions are present in some BRCA1-related breast cancer, they are infrequent and are not recurrent.

Verbatim – from S.O.W

This section refers to the second set of 3 aims

4. Transcriptome profiling of 13 breast cancer cell lines of known paclitaxel sensitivity profile (months 13-16)
5. To search for putative "splicing signature" that correlates with paclitaxel treatment (months 17-19)
6. To validate the putative "splicing signature" for paclitaxel treatment in breast tumor samples (months 20-24)

Unfortunately, due to the departure of Dr. Jun Zhu from Duke University to the NIH, where he was appointed to a non-breast cancer position, we were unable to develop these three themes. Moreover, we were unable to pursue sensitivity to paclitaxel as indicated in these aims. We did, however, have in-house expertise with the tyrosine kinase inhibitor dasatanib (Sprycel®) from our colleague at McGill, Dr. Raquel Aloyz, and therefore we applied our splicing analyses to her two breast cancer cell lines – one resistant and one sensitive to dasatinib. We have identified many promising splice site mutations, indels and SNPs that we still investigating, but overall, it does not look as if we will be able to use these data to find better markers for successful dasatinib treatment of breast cancer patients.

Summary of all work done during the granting period

Thus, although the original the molecular barcode strategy was not implemented (see report of co-PI, Dr Zhu), we have made significant findings in two areas – 1) we have identified over 1400 new splice variants and have shown that they are more prevalent in BRCA1-related breast cancer and 3) we have identified 3 novel fusion proteins. Perhaps most importantly, we have trained two young bioinformaticians (E.Lalonde and K. Ha) who both have bright futures in cancer research.

Key Research Accomplishments

- One published paper directly and entirely funded by the DOD
- One manuscript directly and entirely funded by the DOD is under review
- Development of novel approaches to analyze RNA Seq data
- Identification of >1400 novel splice site junctions in breast cancers and breast cancer cell lines
- Identification of 2 novel fusion proteins in breast cancer cell lines and one novel fusion protein in a breast cancer arising in a BRCA1 mutation carrier

Reportable Outcomes

Posters and Presentations

E. Lalonde

Characterizating the BRCA1-deficient breast cancer transcriptomes by RNA-Seq [presentation]. Era of Hope Meeting, Orlando, Florida, August 4th 2011 (to follow)

Applications of next-generation sequencing in biomedicine: Exome sequencing and RNA-Seq [poster]. Human Genetics Research Day, McGill University, May 31 2011.

Characterization of BRCA1-deficient breast cancer transcriptomes by RNA-Seq reveals novel transcript isoforms [poster]. American Society of Human Genetics Annual Meeting, Washington D.C., November 2-6 2010.

Rapid, genome-wide discovery of coding variants with exome sequencing [poster]. Canadian Human Genetics Conference, Banff AB, April 26-29, 2011.

Published papers (work directly or indirectly supported by this award)

Majewski J, Schwartzentruber J, **Lalonde E**, Montpetit A, and Jabado N (2011). What can exome sequencing do for you? Journal of Medical Genomics, published online July 5th, 2011: doi:10.1136/jmedgenet-2011-100223.

Lalonde E, **Ha KCH**, Wang Z, Bemmo A, Kleinman C, Kwan T, Pastinen T, and **Majewski J** (2011). RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. Genome Research, 21, 545-554.

Lalonde E, Albrecht S, Ha KC, Jacob K, Bolduc N, Polychronakos C, Dechelotte P, **Majewski J** and Jabado N (2010). Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. Human Mutation, 31, 918-923. (while this project is not related to breast cancer, the techniques were acquired during the 3 years of the DOD award)

Anastasio N, Ben-Omran T,Teebi A, **Ha KCH, Lalonde E,** Ali R, Almureikhi M, Kaloustian VMD, Liu J, Rosenblatt DS, **Majewski J** and Majewska L (2010). Identification of the gene responsible for Van Den Ende-Gupta Syndrome. American Journal of Human Genetics, 87, 553-559. (while this project is not related to breast cancer, the techniques were acquired during the 3 years of the DOD award)

Alfares A, Nunez LD, Al-Thihli K, Mitchell J, Melançon S, Anastasio N, **Ha KC, Majewski J**, Rosenblatt DS, Braverman N. Combined malonic and methylmalonic aciduria: exome sequencing reveals mutations in the ACSF3 gene in patients with a non-classic phenotype. J Med Genet. 2011 Jul 23. (while this project is not related to breast cancer, the techniques were acquired during the 3 years of the DOD award)

Shuen AY, **Foulkes WD**. Inherited mutations in breast cancer genes--risk and response. J Mammary Gland Biol Neoplasia. 2011 Apr;16(1):3-15.

Martinez-Marignac VL, Rodrigue A, Davidson D, Couillard M, Al-Moustafa AE, Abramovitz M, **Foulkes WD**, Masson JY, Aloyz R. The effect of a DNA repair gene on cellular invasiveness: XRCC3 over-expression in breast cancer cells. PLoS One. 2011 Jan 24;6(1):e16394.

Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. N Engl J Med. 2010 Nov 11;363(20):1938-48.

Kevin C H Ha, Emilie Lalonde, Lili Li, Luca Cavallone, Rachael Natrajan, Maryou B Lambros, Costas Mitsopoulos, Jarle Hakas, Iwanka Kozarewa, Kerry Fenwick, Chris J Lord, Alan Ashworth, Anne Vincent-Salomon, Mark Basik, Jorge S Reis-Filho, **Jacek Majewski & William D Foulkes**: Identification of gene fusion transcripts by transcriptome sequencing in *BRCA1*-related breast cancers and cell lines. BMC Medical Genomics, 2011, Oct 27;4:75 (Highly accessed, Dec 9 2011)

Submitted papers

Lalonde E, Li L, **Ha KCH**, Cavallone L, Natrajan R, Lambros MB, Mitsopoulos C, Hakas J, Kozarewa I, Fenwick K, Lord CJ, Ashworth A, Vincent-Salomon A, Sapino A, Hardisson D, Reis-Filho JS, Basik M, **Foulkes WD and Majewski J**. Global analysis of previously unrecognized alternative splicing in BRCA1-related breast cancers. Genome Medicine [under review].

NB Kevin Ha was the graduate student of Dr. Majewski and Emilie Lalonde is the graduate student of Dr. Majewski and Foulkes (co-supervision).

Conclusion

In this proposal, we have set out to evaluate the importance of splicing for breast cancer biology. We are employing a number of state-of-the-art technologies – a custom-made capture array to study splicing events in breast cancer and RNA sequencing, which is a relatively unbiased approach to the problem. We hope that our approach will lead to significant insights into the more general question of the importance of alternative splicing in breast cancer biology.

References

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. Genome Res, 12(4), 656-664.

Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., et al. (2009). Transcriptome sequencing to detect gene fusions in cancer. Nature, 458(7234), 97-101.

Prensner, J. R., & Chinnaiyan, A. M. (2009). Oncogenic gene fusions in epithelial carcinomas. Curr Opin Genet Dev, 19(1), 82-91.

Shtivelman, E., Lifshitz, B., Gale, R. P., & Canaani, E. (1985). Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. Nature, 315(6020), 550-554.

Stephens PJ, McBride DJ, Lin ML, et a. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009 Dec 24;462(7276):1005-10.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X., et al. (2005). Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. Science, 310(5748), 644-648.